Pattern Recognition Letters 30 (2009) 1059–1066

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Human action recognition by feature-reduced Gaussian process classification

Hang Zhou [a,*], Liang Wang [b], David Suter [a]

[a] Department of Electrical and Computer Systems Engineering, Monash University, Vic. 3800, Australia
[b] Department of Computer Science and Software Engineering, The University of Melbourne, Vic. 3010, Australia

ARTICLE INFO

Article history:
Available online 1 April 2009

Keywords:
Human action recognition
Characteristic-based descriptor
Gaussian process classification
Spectral feature reduction

ABSTRACT

This paper presents a spectral analysis-based feature-reduced Gaussian Processes (GP) classification approach to recognition of articulated and deformable human actions from image sequences. Using Tensor Subspace Analysis (TSA), space–time human silhouettes extracted from action sequences are transformed to a low dimensional multivariate time series, from which structure-based statistical features are extracted to summarize the action properties. GP classification, based on spectrally reduced features, is then applied to learn and predict action categories. Experimental results on two real-world state-of-the-art datasets show that the GP classification outperforms a Support Vector Machine (SVM). In particular, spectral feature reduction can effectively eliminate the inconsistent features, while leaving performance undiminished. Moreover, compared with Automatic Relevance Determination (ARD), the spectral way for feature reduction is more efficient.

1. Introduction

Human action recognition aims to understand patterns of human movements from video sequences, by classifying actions into known categories such as walking or jumping. This growing interest is strongly driven by a wide range of promising applications such as visual surveillance and human–machine interfaces. However, due to the great variation in the captured human action sequences (with respect to different instances or different persons: with various body types, action styles and speeds), how to extract distinguishable features to describe actions and to accurately model and recognize the actions, still remains a challenge.

'State-space' approaches using probabilistic models such as Hidden Markov Models (HMMs) (Brand et al., 1996; Nguyen et al., 2005) and Conditional Random Fields (CRFs) (Sminchisescu et al., 2005) have been widely used to model human action patterns. However, such probabilistic models are generally of high computational complexity, since detailed statistical modeling is usually required. Moreover, these involve assumptions about the probability distributions of variables of the dynamical models and the consequent development of inference methods as well as model parameter learning algorithms.

In contrast, being a kernel-based non-parametric model, the Gaussian process (GP) (Rasmussen and Williams, 2006) is more computationally tractable. General properties of the GP kernel are controlled by a few hyperparameters which can be estimated under the Bayesian framework. Furthermore, GP is used as a Bayesian prior to express beliefs about the underlying functions being modeled, which is linked to data via the likelihood. The posterior distribution can be directly calculated given the training data.

So far, the solutions applying GP to human action analysis include Wang et al. (2008), Raskin et al. (2007) and Zhou et al. (2008). Gaussian Process Dynamical Models (GPDM) are proposed in (Wang et al., 2008) for non-linear time series analysis with application to human action capture data. It is essentially a 'state-space' solution which models both the distribution of the observed data and the dynamics in the latent space. Raskin et al. (2007) proposed to combine GPDM and annealed particle filtering for tracking and classifying human actions. GP classification for human action recognition was first investigated in (Zhou et al., 2008) with better results than using a Support Vector Machine (SVM).

However, an issue for GP classification (as well as other classifications) is that each of the extracted input features is usually not guaranteed to have positive effect on the classification result. Including more input features could ultimately lead to poor classification outcomes. Only those features that are most relevant or beneficial to the prediction outputs should be selected. Neal (1996) use Automatic Relevance Determination (ARD) to assign each input a weight value (ranging from 0 to 1). The weight values have independent Gaussian prior distributions with standard deviation given by the corresponding hyperparameter with some prior. Then, the posterior distribution of the hyperparameters is calculated given the training data. The values of the hyperparameters are proportional to the corresponding input weight values. An issue for ARD is: what prior should be used for the hyperparameters? (as it will have a significant impact on the accuracy of

* Corresponding author. Tel.: +61 3 99053454.
E-mail addresses: hang.zhou@eng.monash.edu.au (H. Zhou), lwwang@csse.unimelb.edu.au (L. Wang), d.suter@eng.monash.edu.au (D. Suter).

0167-8655/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2009.03.013

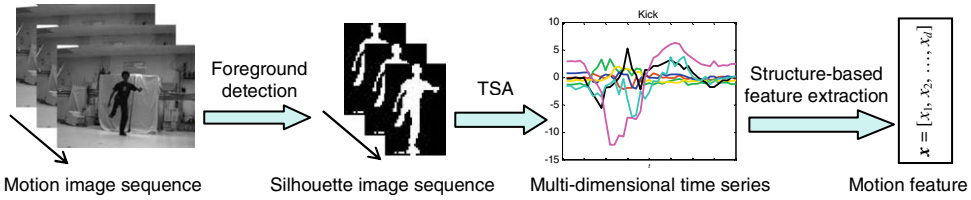

Fig. 1. From action image sequence to characteristic-based descriptor.

estimation results of the hyperparameters). Alternatively, Zhou and Suter apply spectral feature pruning in (Zhou and Suter, 2008). In that approach, the most homogeneous features (with similar 'effective dwell time') are retained and the rest discarded. This was the first time feature selection using spectral techniques has been applied to GPs. In essence, the (actually anisotropic) data have been modified to better match the (isotropic) kernel. However, that still leaves the problem of how to fit better to GP kernel for those data whose features have the same 'effective dwell time'.

To address this problem, this paper uses an additional measure called "major bandwidth".[1] In our approach, characteristic-based feature descriptors (Wang et al., 2008; Wang et al., 2006) are first used to transform time-varying dynamic features of varied-length action sequences to fixed-length feature vectors (and thus the temporal classification problem is converted to a static classification one), which enables the use of GP. Spectral analysis is then applied to the extracted features where the most consistent features with similar "major bandwidth" are chosen as the inputs for GP classification. The "major bandwidth" measure makes it possible to choose the most homogeneous features among those with similar "effective dwell time" as described in (Zhou and Suter, 2008). Compared with ARD, the spectral way for feature selection avoids the need to choose priors for the hyperparameters that exists in ARD (Neal, 1996). Extensive experimental and comparative results on two recent action datasets demonstrate the effectiveness of our approach.

In addition, we have also chosen Nearest-Neighbor (NN) and Support Vector Machine (SVM) to implement baseline experiments. Since NN is one of the simplest and most common machine learning algorithms and SVM is the state-of-the-art kernel-based learning method, they are the appropriate methods to be compared with GP.

The remainder of this paper is structured as follows. Action feature extraction is introduced in Section 2 and a brief review of GP classification is given in Section 3. Section 4 describes the feature reduction algorithm. In Section 5, experimental results are presented and discussed, prior to a summary of main conclusions of the work in Section 6.

2. Feature extraction of actions

Given a database consisting of v action sequences $M = \{M_1, M_2, \ldots, M_n\}$, we extract informative space–time silhouettes to represent the actions performed. The process of feature extraction is illustrated in Fig. 1. For each action sequence M_i including T_i image frames, i.e., $M_i = \{I_1^i, I_2^i, \ldots, I_{T_i}^i\}$, $i = 1, 2, \ldots, n$, the associated sequence of human silhouettes can be obtained by foreground detection techniques. Since the size and position of the foreground region vary among different frames, we construct normalized silhouettes $S_i = \{S_1^i, S_2^i, \ldots, S_{T_i}^i\}$ by centering the silhouette images and normalizing them to the same dimension of $n_1 \times n_2$.

To represent human actions in a more compact subspace rather than in the high dimensional image space, we adopt a structured dimensionality reduction method, i.e., Tensor Subspace Analysis (TSA) (He et al., 2005), to perform subspace learning of the articulated action space. TSA preserves the spatial information of silhouette images by representing an image as a second-order tensor (or a matrix). Given a set of m normalized silhouette images from the training data $\{S_1, S_2, \ldots, S_m\}$ in $R^{n_1} \otimes R^{n_2}$, TSA aims to find two transformation matrices U of size $n_1 \times l_1$ and V of size $n_2 \times l_2$ that map these silhouette images to another set $\{Y_1, Y_2, \ldots, Y_m\}$ in $R^{l_1} \otimes R^{l_2}$ $(l_1 < n_1, l_2 < n_2)$, such that $Y_i = U^T S_i V$. For more details about TSA, the reader may refer to He et al. (2005). In the learned tensor subspace, any silhouette sequence S_i can be accordingly projected into a trajectory $P_i = \{P_1, P_2, \ldots, P_{T_i}\}, P_i \in R^{l_1} \otimes R^{l_2}$. We can regard P_i as a form of multivariate time series with the number of dimensions $l = l_1 \times l_2$.

Structure-based statistical features are then extracted to summarize the multivariate time series: which turns action time series of different lengths into a feature vector of the same length. The nine most informative, representative and easily measurable characteristics are chosen to summarize the time series structure (Wang et al., 2006): *trend*, *seasonality*, *serial correlation*, *non-linearity*, *skewness*, *kurtosis*, *self-similarity*, *chaotic*, and *periodicity*. Based on these characteristics, corresponding metrics are calculated to form the structure-based feature vectors (Wang et al., 2008), called characteristic-based descriptors. For each dimension of P_i, we obtain 13 statistical features. Thus, a multi-dimensional time series P_i is summarized by a δ-dimensional ($d = 13 \times l$) feature vector $\boldsymbol{x}$.

3. GP classification

A Gaussian processes (GP) is a collection of random variables, any finite number of which has a joint Gaussian distribution (Rasmussen and Williams, 2006). A GP is fully specified by its mean function $m(\boldsymbol{x})$ and kernel function $k(\boldsymbol{x}, \boldsymbol{x}')$, expressed as:

$$f \sim GP(m, k). \tag{3.1}$$

The GP classification process models the posterior directly. The GP prior is represented by the kernel function which characterizes correlations between points in the training data (which is a sample process). The kernel function's hyperparameters can be learned from the training data.

The kernel function used in this paper is the Radial Basis Function (RBF), also called the Squared Exponential (SE) function or Gaussian function (Snelson, 2006), i.e.,

$$k_{\mathrm{RBF}}(x - x') = \sigma_0^2 \exp\left[-\frac{1}{2}\left(\frac{x - x'}{l}\right)^2\right], \tag{3.2}$$

where x and x' are input pairs, l is the characteristic length scale and σ_0 is the signal variance. RBF is an isotropic kernel whose main advantage over non-isotropic versions is the simplicity. RBF is shift and rotation invariant on both signal and frequency domains as shown in Fig. 2. The isotropy of the RBF kernel is the fundamental property for feature reduction (see Section 4).

[1] This paper is actually an extended version of our previous work described in (Zhou et al., 2008).